在前期搜索过程的基础上，对搜索的细节和可视化作了进一步的细化，并重新思考了文章的创新点和实现方法。

论文整体思路：

1. 搜索

前期已经介绍了搜索到具体方法，但是一直没有很明确的参考文献来支撑我的思路。SIGMOD’08上清华李国良提出了r半径Steiner图搜索思想，对非结构化、半结构化和结构化数据可以执行统一的关键字搜索。该论文的引用率达到了201次，后续很多人在这种思想的基础上开展了很多工作。在论文中提到了通用搜索引擎难以解决的一类问题：关键字跨页面。比如说，当搜索2008年在加拿大举办的“数据集成”会议时，可能输入关键字“Conference 2008 Canada Data Integration”。SIGMOD 2008是在加拿大召开的，而且数据集成是其中的一个主要研究议题。但是在实际搜索时，SIGMOD 2008不仅没有出现在Top 10中，甚至前100个结果中都没有SIGMOD 2008。其主要原因是因为”2008，Conference”和”Data Integration”并没有出现在同一页面中，搜索引擎在计算相似性时其值没有其他结果高。

对于微博这一类短文本数据，同样具有这样的特点，很多关键字并不出现在同一条消息，但是观察某个用户某段时间内发布的消息，却涵盖了这些关键字或类似的关键字。因此可以借助于类似的r半径Steiner图搜索思想来完成关键字搜索。

但是在李国良的论文中同样提到了另外一个问题。r半径Steiner图搜索是在一定范围内寻找直接或间接的关键字，这样带来一个很明显的结果是：某些关键字并不是用户所关心的，根据这些关键字查询的结果并非用户所期待的。李国良的解决办法仍然是借助于数据库中top-k问题的思路，计算ranking。但是如果将查询结果进行图形化表示，并提供适当的交互手段供用户选择，能更直接地解决这类问题。



在实现上，根据我目前文本分析的结果可以将模型表示成上述结构。根据每个用户发布的消息，可以抽取Entity，每个Entity都有相关的Attribute。当输入关键词时，根据属性得到Entity，再根据Entity的每个Attribute再去执行完整的搜索。这种实现与r半径Steiner略有不同。

1. 可视化

先前所考虑的可视化方法一直陷于如何表达多个关键字（Entity）及查询结果分类表示上，一度陷入无法合理表示的境地。实际上，由于可视化的对象是用户，而不是微博消息本身，所以只要能够将查询结果进行合理分类即可。至于实体之间的关系并不具有很实际的意义，但是可以作为相似性对结果进行布局。

之前考虑过用Venn图，但是由于查询结果可能比较多可能会造成视觉混乱，而且无法表示相似性。因而，改用语义Bubbleset来进行表达。也就是说bubbleset中的节点按照Entity来进行布局，而超图的每条边代表了某个主题下的搜索结果。

1. 交互

交互仍然考虑使用Dynamic Search

1. 最终结果

最终结果一方面显示关注度较高的用户，另一方面将按照消息的回复和转发来进行排序，不考虑相似性。